Data Governance – Simplifying Machine Learning Model Deployment

Abstract

Although machine learning (ML) offers many advantages when it comes to predicting outcomes and identifying with novel insights, deploying ML models in itself is a challenging task. Apart from assembling the various constituent components of its architecture, coding teams are also required to ensure the integrity and quality of data used for operationalizing the models – a task which is made more complex without strong data governance, especially if an ML model is to be integrated with third-party data sources.

This paper discusses the foundational elements required for setting up such a governance framework along with relevant details aimed at simplifying ML model deployment.
Machine Learning in Action

What if you were running a transportation business and wanted to reduce the cost of maintaining engines? Machine learning (ML) techniques, in that regard, can be a game changer. If deployed correctly, it can help you predict just when a bus or any other vehicle in your fleet needs to be recalled for proactive maintenance. In turn, this will help you minimize the impact of general wear and tear, diminish failures in the field and the need for a major overhaul, and keep your fleet on the road for as long as possible so that you don’t miss out on revenue opportunities.

While developing such an ML model, you might soon come to realize that factors which influence engine use and performance are myriad. These can range from the individual engine's behavior to weather patterns, the number of passengers ferried daily, on-road traffic volumes, and festivals and events on travel routes to name just a few. Consequently, you will realize that although you might have complete access to some of this data, acquiring information such as weather and traffic volume forecasts means having to tap into external data sources.

Initially, you will begin by just designing and developing the basic ML architecture. It can comprise a multi-layer neural network (NN), a convolutional neural network (CNN) or even a recurring neural network (RNN) with additional functionality layers on top of it. While building the model, other considerations such as incoming data volumes, expected performance, and the time required for training the model will have to be taken into account. You will also need to spend a considerable amount of time validating, normalizing, and ensuring that the data is labeled properly for training the structured algorithms. Thereafter, you will have to optimize the algorithm, reduce bias and variance errors, and define target values before you finally have a model with the desired level of accuracy which is ready to be deployed in a production environment. This sets the stage for the next stage of challenges that you will have to overcome.
Outputs from Training Activities

Along with the ML model, a key set of outputs that can be acquired post training activities constitutes an inventory of data sources along with metadata pertaining to volume, veracity and velocity description, technical format, physical locations, as well as the access and control descriptions of the information streams. Using them, you will be able to develop a number of normalization, transformation, and classification rules for shaping the data into a consistent and coherent format based on which you have trained your ML model.

Operational Architecture

In light of all of these new outputs, the ML model’s operational architecture has to be flexible enough to integrate discrete data components. In case of the transportation business for example, this can include daily engine metrics, real-time passenger count as they board and disembark from the vehicles, traffic conditions, and daily weather forecasts on a low-volume monthly basis.

Each of these sources have to be organized according to how big the associated datasets are, how often they are generated, and how they are validated. There must be policies in place for controlling and managing each distinct data stream. And, as is the case with any data source, you will need the logical and physical data models. For many companies, this may as well be the first time they are using external data sources as a key tool to streamline their decision-making process.

Data Quality Strategy

To make sure that the ML model is functioning as intended, maintaining data quality will be an imperative. To do so, data consistency, relevance, and overall accuracy will need to be monitored while assessments and metrics for measuring how well data sources are performing should ideally be established. At this point, quality gaps or bugs in the data sources such as missing or unavailable data and the effort required for cleansing can become a significant burden. In turn, this can render a data source nearly unusable and may substantially affect the final trained algorithms.
While working with external data sources, you might face a high latency problem which can make it difficult for the ML model to derive predictive insights in near real-time. To overcome this, you will need to determine how much additional computational power is required to run potentially complex algorithms on demand.

**Implementation Roadmap**

Apart from the basics of the ML model’s architecture that has been discussed here already, data ingestion, transformation, normalization, and loading capabilities for multiple sources needs to be integrated.

Despite the test results gathered from the training stage, the ML model’s functionality will still need to be assessed in terms of the quality of its output. In terms of the maintenance prediction model we have been talking about, when it suggests that a specific vehicle needs to be taken off the road, the recommendation will need to be cross-validated for its accuracy by mechanics and engineers. If they disagree with the prediction and concur that the vehicle could have efficiently operated for another month, the feedback can be used to improve the ML model further. The overall goal at this stage is to establish a clear evaluation process to collect feedback on performance and validate them in line with acceptable tolerance levels. Ideally, there should be strong governance in place to keep this process operating continuously.

A further extension to this is the concept of A/B testing. As ML engineers design new models, their operations teams will need to deploy mechanisms for running comparative tests. There should also be a feature which enables it to pass the same data subsets through individual models for generating automated comparisons.

Over time, the ML model will need to be retrained, not only if it produces results which deviate too far from normal expectations but also as part of general maintenance procedures. It could also be because better data sources have recently opened up or new features need to be added to the existing information streams or perhaps because a bias has been detected. As a corrective measure, new ML models can be released much like any other software update. If the deployment approach comprises a series of ML models where
the output of one flows into another (as is the case with the maintenance prediction example), a clear framework and a set of metrics for measuring each model's performance need to be established.

An Agile Approach

There is a clear need for continuous integration and deployment tools, especially because the terms of use pertaining to different datasets often vary while new models and data sources continue to emerge. If organizations wish to rapidly capitalize on the benefits they offer, an agile infrastructure will be essential, especially when models are not performing as they had in the lab. This creates a true biz-dev-ops environment where every stakeholder including data scientists, developers, and operations teams focus on diagnosing, testing, and deploying potential high impact models.

Strong policies for version control on both the model and its parameters will have to be managed and maintained. For example, a new set of parameters may be a minor model release, whereas a complete new model may be a major release.

Data Governance

The implementation of these policies with respect to governance is the next step. Most organizations have established data management rules for internal data sources. However, in scenarios where companies want to use external sources and third-party data brokers, clear control and quality metrics will have to be established. While using external sources, any associated procurement policy and purchasing rights will have to be put in place. Furthermore, specific governance regarding associated intellectual property rights and usage rights will have to be incorporated into the framework, as well as adherence to the General Data Protection Regulation (GDPR) regulations.
The Way Forward

Developing and training ML models is challenging to say the least. Despite the numerous frameworks and toolkits that developers have access to today, making an ML model fully operational poses challenges that are unlike those associated with general software development. Apart from the skills and capabilities required, an ML development project should take note of:

- The impact it will have on the operational architecture
- External data sources and how they will be used on a day-on-day basis
- Quality of output, how it can be measured, and what can be done to improve it
- Operations monitoring and validation

As organizations become increasingly data-driven, they will come to depend on AI and ML as the primary source for business intelligence. Depending on how extensive the practice grows, it will begin to exhibit a number of complex operational challenges. Imagine the transport business running ML models for not just evaluating engine health but also for measuring supplier performance, facilitating partner management, and driving sales and marketing campaigns. As such, at any given time, there could be hundreds of ML models in operations – each with staggering data requirements and computational needs. Governance and change management, in such a scenario, will be the key for deriving the maximum value from operations.
About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com